Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2020 ; 17(3): 868–876. doi:10.1109/TCBB.2018.2869738.

Preprocessing Sequence Coverage Data for More Precise Detection of Copy Number Variations

Fatima Zare[*], **Sardar Ansari**[†,§], **Kayvan Najarian**[†,‡,§,¶], **Sheida Nabavi**[*]

[*]Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut, 06269

[‡]Department of Emergency Medicine

[§]Department of Computational Medicine and Bioinformatics

[¶]Department of Electrical Engineering and Computer Science

[†]University of Michigan, Ann Arbor, Michigan, 48109

Abstract

Copy number variation (CNV) is a type of genomic/genetic variation that plays an important role in phenotypic diversity, evolution, and disease susceptibility. Next generation sequencing (NGS) technologies have created an opportunity for more accurate detection of CNVs with higher resolution. However, efficient and precise detection of CNVs remains challenging due to high levels of noise and biases, data heterogeneity and the “big data” nature of NGS data. Sequence coverage (readcount) data are mostly used for detecting CNVs, specially for whole exome sequencing data. Readcount data are contaminated with several types of biases and noise that hinder accurate detection of CNVs. In this work, we introduce a novel preprocessing pipeline for reducing noise and biases to improve the detection accuracy of CNVs in heterogeneous NGS data, such as cancer whole exome sequencing data. We have employed several normalization methods to reduce readcount’s biases that are due to GC content of reads, read alignment problems, and sample impurity. We have also developed a novel efficient and effective smoothing approach based on Taut String to reduce noise and increase CNV detection power. Using simulated and real data we showed that employing the proposed preprocessing pipeline significantly improves the accuracy of CNV detection.

Keywords

Copy Number Variation; Whole-Exome Sequencing; Signal Processing; Taut String; Cancer; Normalization; Denoising

I. INTRODUCTION

Copy number variations (CNVs) are a critical and common source of variation in the human genome. A deletion or amplification of a segment of a genome (ranging from a few hundred

fatemeh.zare@uconn.edu.

base pairs to a few mega base pairs) is defined as a CNV; and compared to single-nucleotide polymorphisms (SNP), CNVs incorporate a greater proportion of the genome (4.8–9.5% of the genome) [1]. Recently CNV has gained considerable interest as a type of genomic variation and evidences have shown associations between CNVs and many diseases including cancer [2]. With the arrival of next generation sequencing technologies and due to importance of detecting CNVs, many CNV detection tools have been developed; however it has been shown that the agreement among them on detected CNV segments is low and they generate false positives [3]. One of the reasons that CNV detection algorithms detect false CNV segments is due to noise and biases of sequencing data. Employing effective and efficient denoising and normalization methods can significantly improve the detection accuracy of CNV detection algorithms. The main aim of this study is developing an effective preprocessing method to reduce noise and biases for better detection of CNVs, especially in cancer, using next generation sequencing (NGS) data.

NGS is a popular strategy for genotyping and by generating hundreds of millions of short reads in a single run can provide a comprehensive characterization of CNVs [4]. NGS attempts to consolidate the advantages of Sanger sequencing and SNP array technologies. The greatest benefit of NGS over conventional Sanger sequencing is the capacity to sequence a huge number of reads in a single run at a low cost [5]. Nonetheless, due to short read lengths and the complexity of the genome, NGS technologies have introduced many new challenges for the analysis of CNVs [6]. In addition, detection of CNVs from NGS data is relatively new and there is no commonly agreed method for CNV detection.

Recently, whole-genome sequencing (WGS) and whole-exome sequencing (WES) have become primary strategies for NGS technologies in CNV detection and for studying human diseases. Several studies have suggested to use WGS data for CNV detection. Since WGS is considered too expensive for research involving large cohorts, cost-effective WES has become the primary strategy for sequencing, especially in biomedical research such as cancer studies. Because of the abundance and popularity of WES data for detecting clinically relevant aberrations in cancer, in this study we focus on WES data. WES data has several technical issues. Unlike WGS, WES requires PCR amplification that leads to more GC bias. Also, hybridization in WES causes low or no coverage in some regions of the genome, which introduces more mappability bias. These issues need to be considered for an appropriate CNV detection method for WES data [7].

There are several approaches for detecting CNVs using NGS data [8]. Among them, the read depth (RD) approach is the most applicable method especially for WES data. The RD based approaches assume that the density of short reads is locally proportional to the copy number [9]. In the RD based approach, mostly a non-overlapping sliding window is used to count the number of short reads that have overlap with the window. These readcount values are used to identify CNV regions [3]. The correlation between the readcount value and copy number of a specific genomic region is the primary idea behind the RD based CNV detection methods. However, existence of noise and biases distorts the relationship between the readcounts and copy numbers that introduces challenges for CNV detection. As a result, for having a precise CNV detection, these biases and noise should be removed from readcount data before detecting CNVs.

In general, CNV detection methods consist of two major parts: preprocessing and segmentation. The aim of the preprocessing part is reducing noise and biases for better identification of CNV regions in the segmentation part. The focus of this work is on the preprocessing part. There are several types of noise and biases for NGS data: GC bias, mappability bias, sample contamination, sequencing noise and experimental noise.

GC bias has been introduced for the first time in [10]. It has been observed that regions with low or high GC content have low readcounts compared to other regions. In fact, there is a unimodal relationship between readcounts and G and C bases in a genome [11]–[14]. GC bias is neither linear nor consistent among different samples. Several methods have been proposed to model and remove GS bias [15], [16], [17]. The most popular approach for removing GC bias is the Loess regression method [13], [15], [18]. The Loess regression method removes reads from regions with very high coverage compared to the expected value of coverages and will add reads to regions where very few reads are observed.

Furthermore, a huge number of NGS reads cannot be uniquely mapped to the reference genome due to short length of reads and the presence of repetitive regions within the reference genome. Mutations and sequencing errors can lead to incorrect mapping of short reads as well. These errors cause ambiguities in the alignment process, resulting in mappability biases [19]. To remove mappability biases in cancer data, CNV detection methods mostly use the number of uniquely mapped short reads in tumor and matched normal samples and apply a Loess regression method [20], [21].

A very challenging problem in CNV detection is detecting focal (narrow) CNV regions under extreme noise [22], [23]. A few CNV detection tools have employed denoising methods such as the discrete wavelet transform (DWT) [24] and Bayesian approaches [25], [26] for noise cancelation. Noise can corrupt readcount data, which can be seen as readcount signal, and signal-to-noise ratios heavily influence the accuracy of CNV detection. If the level of noise in readcount signal is high, CNV detection algorithms are likely to detect many false positives and miss focal aberrations. Signal processing techniques, which have been long used for effective noise cancellation, can be extremely useful for improving CNV detection by identifying and removing noise from the CNV readcount signals [27]–[29].

Complexity of tumor samples imposes another challenge to CNV detection. Tumor samples are heterogeneous and contaminated by normal cells. In other words, sequencing tumor samples provides reads from the admixture of normal and sub-clonal cancer cells [18]. Tumor contamination has been evaluated by visual examination of tumor cells by a pathologist or through image processing [30]. Recently, some computational methods have been introduced to estimate tumor contamination and to use it for normalizing readcount data in CNV detection [18], [31]–[33].

In this work, we introduce a new pipeline for preprocessing readcount data in cancer data for detecting CNVs more accurately. The pipeline includes filtering outlier readcounts, removing GC and mappability biases, reducing noise, and normalizing for tumor purity. Due to the important role of noise cancellation in improving CNV detection power, we have

developed a new efficient and precise denoising method based on a signal processing technique, Taut String.

II. METHOD

The proposed preprocessing pipeline can be divided into 5 blocks: 1) Filtering outlier readcount data; 2) Removing GC bias from both tumor and normal readcount data; 3) Calculating the ratios of tumor and normal readcounts for each genomic window and removing mappability bias; 4) Eliminating noise from the normalized readcount data; and 5) Eliminating the effects of the tumor contamination by normalizing the denoised ratios. The outputs of the preprocessing pipeline then input a segmentation algorithm for the detection of CNV regions (Figure 1). We used the circular binary segmentation (CBS) method [34] for segmentation. We will explain the details about the preprocessing pipeline in the following sections.

A. Filtering outlier Readcounts

A sliding window approach is used to compute the GC% and readcount value for each genomic window [35]. The size of window is optional. We used a window of size 100 in this work. Windows with readcounts and GC content in the bottom and top 1% quantile are considered as outlier windows and are removed from the data.

B. Removing GC Bias

We employ the weighted Loess regression method for removing GC bias [15]. In this method, a regression analysis is applied to the mean of readcount values that are from windows with specific GC content. For both tumor and normal samples, the mean of readcount values with specific GC content, m_{gc}, is computed for each possible percentage of GC content as:

$$m_{gc} = \frac{\sum_{i=1}^{n_{gc}} d_i^{gc}}{n_{gc}}, \tag{1}$$

where gc is the percentage of the GC content, n_{gc} is the number of windows that have gc, and d_i^{gc} is the readcount for window i that has GC content of gc. This method uses a weight for each GC content that is equal to the number of windows with the corresponding GC content, $w_{gc} = n_{gc}$. Insufficient read coverage of some percentile of GC content can lead to local extremes. This means that if there are a few windows with GC content of gc (low n_{gc}), then their corresponding m_{gc} value would be significantly higher or lower than m_{gc} values that are corresponding to many other windows (high n_{gc}). These local extremes are removed by the Loess regression method, which will lead to smoother values of m_{gc}. Then, a weighted Loess regression is applied to all m_{gc} values with their corresponding weights. Finally, the number of reads for each window is corrected through the Equation (2):

$$\hat{d} = d - (d_{loess} - av(d)), \tag{2}$$

where d is the readcount value of a particular window before applying the Loess correction, $dloess$ is the windows' smoothed readcounts after the Loess correction, and $av(d)$ is the mean of readcounts for all windows.

C. Removing Mappability Bias

The ambiguities in alignment can result in mappability bias in RD based CNV detection methods [36]. To eliminate this bias, we employ the method introduced in [20]. In this method, the log2 ratio of tumor and normal readcounts for window i, r_i, computed as:

$$r_i = \log_2\left(\frac{\hat{d}_{T_i}}{\hat{d}_{N_i}}\right), \tag{3}$$

Where $\hat{d}_{T_i}$ and $\hat{d}_{N_i}$ are readcounts for window i in the tumor and normal genomes, respectively, after GC bias correction. Then, we use the number of uniquely mapped bases in tumor and normal (I_T and I_N) to remove the mappability bias from r for each window i:

$$r'_i = r_i \cdot \left(\frac{I_N}{I_T}\right). \tag{4}$$

I_N and I_T are obtained using Equation (5), where d_M is the number of mapped reads, D is the duplicated mapped reads and $av(L)$ is the average length of reads, obtained via SAMtools software package [37].

$$I = d_M \cdot \left(1 - \frac{D}{d_M}\right) \cdot av(L)\,. \tag{5}$$

D. Noise Cancellation with Taut String

Accuracy of CNV detection is significantly affected by the noisiness of the readcount data (signal). CNV detection methods identify many false CNVs (false positives (FPs) and false negatives (FNs)) when apply to noisy readcount data. As a result, removing noise is a critical issue in CNV detection. The log2 ratios of readcounts (after bias cancellation, Equation (5)), for each genomic window i on the genome, can be model as Equation (6).

$$r'_i = f_i + y_i, \tag{6}$$

where y_i is independent and identically distributed (iid) noise, drawn from a normal distribution $N(0, \sigma_N^2)$ with mean of 0 and standard deviation of σ_N. The goal is to recover original signal f from the noisy observed signal r′.

Selecting an appropriate noise cancelation method depends on the characteristics of the noise and signal. From a signal processing point of view, readcount data are sparse, discrete, and piecewise constant. There are several techniques for noise cancelation. Kernel estimators and Fourier based filtering methods [38], [39] are popular approaches for noise

reduction. However, these methods cannot perform well when denoised signals have several amplitudes such as CNV segment data. As a result, they can reduce noise, but they are not able to preserve edges.

Another drawback occurs when the noise and signal Fourier spectra overlap. In this situation, these linear approaches cannot separate spectra correctly [40]. In addition, detection of small CNV segments in a noisy environment is another challenge. Small CNV segments are mostly discarded through linear filtering approaches. In general in readcount data, amplitude distortion is more than spectra location distortion by noise. Non-linear methods that consider amplitudes rather than locations of the spectra in their noise removing procedure can work better in this situation. Also, in addition to protect narrow CNVs while removing noise, accurate detection of breakpoints of the CNV segments is very important and we need to use a noise reduction method that preserves edges.

Sparse representation of signals has been used for a wide range of applications including removing noise. Discrete wavelet transform (DWT) is type of a linear transformation which are used for obtaining the sparse representation. DWT has a low computational complexity compared Fourier transform. However, DWT has several drawbacks such as oscillations, shift variance, aliasing and lack of directionality [41].

It has been shown that the solutions of the total variation (TV) regularization are sparse and can remove noise from signal. TV based regularization method has been widely used in the signal processing community to remove noise from signals while preserve edges and small local changes. In this paper, we use an algorithm based on the Taut String method that is known for providing extremely efficient solutions to 1D-TV problem in O(N). Taut String, introduced in [42], is an efficient and effective non-linear denoising method [43], [44] and it can solve a penalized least squares functional with considering total variation norm based penalty [44], [45].

Taut String is a nonparametric smoothing method, which has the ability to detect local extreme values in a very noisy data. We use ϵ to define the level of noise of y: $\epsilon = \|y\|_\infty$.

For a fixed $\epsilon > 0$, the goal is to find a unique log2 ratio values, f, such that:

$$\|\mathbf{r}' - \mathbf{f}\|_\infty = \max_i |r'_i - f_i| \leq \epsilon\,. \tag{7}$$

To satisfy Equation (7), f should minimize Equation (8) (norm 2 of a distance operation, DT) and Equation (9) (norm 1 of the second derivative of f) as the optimization objectives:

$$\|DT(\mathbf{f})\|_2 = \sqrt{\sum_{i=1}^{n-1} (f_{i+1} - f_i)^2}, \tag{8}$$

$$\|DT^*DT(\mathbf{f})\|_1 = |f_2 - f_1| + \sum_{i=2}^{n-1} |f_{i-1} - 2f_i + f_{i+1}| + |f_n - f_{n-1}|, \tag{9}$$

Where, DT^*: $R^{(n-1)} \rightarrow R^n$ is dual to DT: $R^n \rightarrow R^{(n-1)}$:

$$DT(f) = (f_2 - f_1, f_3 - f_2, \ldots, f_n - f_{n-1}), \tag{10}$$

and

$$DT^*(b_1, b_2, \ldots, b_{n-1}) = (-b_1, b_1 - b_2, b_2 - b_3, \ldots, b_{n-2} - b_{n-1}, b_{n-1}). \tag{11}$$

By using a linear regression, an estimate of f $\hat{\mathbf{f}}$. , can be computed. In fact, it can be shown that $\hat{\mathbf{f}}$ can be represented as a string between r′ − ϵ and r′ + ϵ that is pulled tight (Figure 2a). $\hat{\mathbf{f}}$ can be computed efficiently in linear time complexity ($O(n)$) [44]. This approach eliminates very lowfrequency noise while keeps the location of breakpoints. The only challenge is obtaining an optimum ϵ, which we used a 10-fold cross validation algorithm to obtain it.

E. Normalizing Tumor Contamination

The tumor samples are admixture of normal and cancerous cells. Distribution of denoised copy number ratios f′(ratios of tumor to normal readcount values ($\mathbf{f}' = 2^{\hat{\mathbf{f}}}$)) can be represented as a mixture normal distribution:

$$pdf(\mathbf{f}') = \sum_{m=1}^{M} a_m N\left(\mu_m, \sigma_m^2\right), \tag{12}$$

where a_ms are the mixing proportions and their values are between 0 and 1. Each μ_m shows a value of the ratio of tumor to normal copy numbers. These numbers can take any value in the set of {0, 1, 1.5, }. We estimate the parameters this normal distribution through the Expectation Maximization algorithm. Using $v = arg\max_m a_m$, we define $\delta = \frac{1}{\mu_v}$.We utilize the tumor contamination model which is proposed by [18] to obtain contamination free readcount ratios c from the denoised readcount ratios f′, considering the contamination proportion λ in tumor samples:

$$\mathbf{c} = 1 + (\mathbf{f}' - 1) \cdot \frac{1}{1 - \lambda}, \tag{13}$$

and

$$\lambda = \frac{1}{M-1} \sum_m (1 - \frac{|\hat{\mu}_m - \hat{\mu}_v|}{\hat{\mu}_v} \cdot \frac{1}{0.5 \times |2(f'_m - f'_v)|}), \tag{14}$$

where f'_m and f'_v are normalized values of $\mu_m$and μ_n and f'_m and f'_v are their corresponding ratios.

III. DATA SETS

To evaluate the performance of the proposed preprocessing pipeline, we have used three sets of data sets: 1) simulated readcount data, 2) simulated sequencing data, and 3) real data.

A. Simulated readcount data sets

We used simulated data to investigate the power of the denoising step in detecting true CNVs and their breakpoints. We generated 10 simulated readcount data sets with known CNVs. To generate these data sets, we used the detected CNV segments from chromosome 1 of real data, obtained by applying Varscan2 [46] and CBS segmentation [47]. The known CNV segments were used for benchmarking. Sampling from the CNV segments of 10 real data sets at 100 bp genomic distances, we generated 10 sets of noiseless readcount signals. Then, we added white Gaussian noise to the generated readcount signals to simulate noisy readcount signals. These simulated data sets do not reflect biases and we used them to evaluate the performance of the denoising block of the pipeline. By adding different levels of noise, we simulated noisy readcount signals with several signal to noise ratios (SNRs) for each of the noiseless generated readcount signals, where the power of noise is σ_N^2.

B. Simulated sequencing data sets

We have also used a CNV simulator, called CNV-Sim (https://github.com/NabaviLab/CNV-Sim) to evaluate the performance of the denoising block. CNV-Sim is a simulation software tool that is highly optimized to make use of existing short read simulators. CNV-Sim gets the reference genome in FASTA format and sequencing targets (exons in the case of WES) in BED format as its inputs. Based on the simulator parameters, a list of CNV regions that are affected by amplifications or deletions is randomly generated. The CNV simulator generates three outputs: (i) a list file that contains the synthesized amplifications and deletions in txt format, (ii) short reads with no CNVs as control in FASTQ format, and (iii) short reads with synthesized CNV as case in FASTQ format. We generated 10 datasets using CNV-Sim for chromosome 1. We used BWA tool [48] to align short reads to the reference genome (hg19) and generated BAM files. Then, using bedtools [49] and 100bp sliding window, we generated readcount data for these simulated sequencing data. These simulated data with known aberrant regions were used to evaluate the performance of the CNV detection tools in terms of sensitivity and specificity.

C. Real data sets

In this study, we used 10 pairs of breast cancer tumor and matched normal WES data sets, provided by the cancer genome atlas (TCGA), to evaluate the performance of the proposed preprocessing pipeline in terms of sensitivity and specificity of detecting true CNVs. We downloaded raw WES data in FASTQ format from the Cancer Genomic Hub (https://cghub.ucsc.edu/index.html). We used BWA software tool [48] to align short reads to the reference genome (hg19) and generated BAM files. Then, by using bedtools [49] and 100bp sliding window we generated readcount data for these samples. We used the CNV results of these samples from the SNP array platform, provided by TCGA, as the benchmark.

IV. RESULTS AND DISCUSSION

A. Results using simulated readcount data

In this section, by using the simulated data, we evaluated the sensitivity of CNV detection and the accuracy of detected breakpoints with and without employing denoising. We compared the performance of the proposed denoising methods, Taut String, with DWT and moving average (MA) denoising methods.

1) Breakpoint Accuracy: Figure 3 shows the simulated readcount signal before and after applying denoising on a noisy signal with σ_N^2 = 0.031 (SNR=7). From Figure 3, It can be seen that Taut String outperforms DWT and MA in preserving edges.

We applied DWT, Taut String and MA denoising methods to the 10 noisy simulated readcount data at different levels of noise with SNRs ranging from 2 to 10 (σ_N^2 ranging from 0.101 to 0.016). Then, we used CBS to detect CNVs' segments from the denoised readcount data. In this analysis, breakpoint accuracy is defined as the percentage of the number of times the start and end points of detected CNVs' segments are exactly the same as in the known CNVs' segments. Figure 4 shows the effects of employing DWT, Taut String and MA on the accuracy of detected breakpoints. As can be seen in Figure 4, using an appropriate smoothing method before segmentation increases the breakpoint accuracy.

DWT and Taut String perform better than MA, especially at higher levels of noise. At higher levels of noise DWT and Taut String methods perform almost similar; but for lower levels of noise Taut String outperforms DWT. The reason for superior performance of Taut String denoising is that it is more powerful to preserve edges. Denoised signals by DWT and MA show more fluctuations at the breakpoints compared to Taut String, which cause less accurate CNV breakpoint detection.

2) Sensitivity of detecting CNV segments: Using the 10 sets of simulate readcount data with several levels of noise, we compared the performance of DWT, Taut String and MA in term of sensitivity in detecting CNV segments. We used segment-based comparison to evaluate the performance of the denoising methods. We used GenomicRanges R package from Bioconductor [50] to obtain overlapping regions between detected CNVs and benchmark CNVs. If a detected amplified/deleted segment has an overlap of 80% or more with a benchmark amplified/deleted segment, it was considered as True Positive (TP). An amplified/deleted segment in the benchmark that does not have an overlap of 80% or more with any detected amplified/deleted regions was called FN. We calculated sensitivity as:

$$Sensitivity = TP/(FN+TP). \quad (15)$$

Figure 5 shows the results of sensitivity analysis for amplified and deleted segments, with thr=±0.2. As expected, all three denoising methods improve sensitivity of CNV detection. However, edge protecting methods (DWT and Taut String) significantly outperform MA. Also, due to better performance of Taut String on preserving edges and protecting narrow

changes in data, it improves the sensitivity of CNV detection slightly better than DWT, especially at lower levels of noise.

B. Results using simulated sequencing data

To evaluate the performance of each block of the proposed pipeline, we computed sensitivity, false discovery rate (FDR) and specificity of CNV detection before and after applying each preprocessing steps. Table I shows definitions for sensitivity, FDR, and specificity. To call TPs, TNs, FPs, and FNs we used a gene-based approach [3], where we first annotated the detected CNV segments to obtain CNV gene lists. We used the CBS method from DNACopy Bioconductor package [47] to detect CNV segments. We used cghMCR R package from Bioconductor [51] to identify CNV genes using Refseq gene identifications. Thresholds of ±0.2 for log2 ratios were used for calling CNV genes. Table II shows the performance of each preprocessing steps using the simulated WES data. GC bias, mappability bias and tumor contamination were not modeled in the simulation and the simulated data do not contain these biases. Therefor, we did not consider GC and mappability bias removing in our analysis for simulated data and we did not include them in Table II. From Table II, we can see that the denoising block improves the performance of the CNV detection.

C. Results using real data

In this section to evaluate the effectiveness of the preprocessing pipeline on detecting true CNVs, we compared the results of CNV detection on real data sets with and without employing the proposed preprocessing pipeline. We also compared the performance of the proposed pipeline with that of the VarScan2 pipeline in terms of sensitivity and specificity in detecting CNVs.

1) Performance of preprocessing blocks: To evaluate the performance of each block of the proposed pipeline, using real data, we used gene-based approach to calculate sensitivity, FDR and specificity of CNV detection before and after applying each preprocessing steps.

As depicted in Table III, each preprocessing block improves the performance of the CNV detection. In overall, by using the proposed preprocessing steps, the sensitivity of detecting amplifications improves from 50.99% to 72.75% and the sensitivity of detecting deletions improves from 60.37% to 84.30%. The performance of the CNV detection method is mostly affected by the denoising block. The denoising block improves the performance of detection for amplifications from 58.57% to 70.15% and from 68.74% to 80.04% for deletions. We also compared the performance of DWT and Taut String methods on real data. As can be seen in Table III, Taut String outperforms DWT in denoising real readcount data and providing higher sensitivity and specificity in CNV detection.

2) Performance comparison with VarScan2: VarScan2 is a preprocessing pipeline that generates normalized readcount data of tumor-normal pairs. The output of VarScan2 inputs a segmentation method for identifying CNV segments. For segmentation, we used the CBS method from the DNACopy Bioconductor package [47]. Table IV shows that the

overall performance of the proposed method is better than VarScan2. The main reason would be using the Taut String method to remove noise from normalized readcount data.

D. Runtime comparison

In this section, we compared the overall runtime of DWT and Taut String methods using real and simulated data sets on a 64-bit Windows 10 Operating System, having 16 GB DDR4 memory and intel core i7–7500U 2.7 GHz CPU. Taut String is linear in time and has the time complexity of $O(n)$. DWT has the time complexity of $O(nlogn)$ [52]. Taut String shows shorter runtime compared to DWT. Using the real datasets, on average, DWT took 30.73 seconds while Taut String took only 5.51 seconds. We observed similar behavior using simulated data. On average, DWT took 26.93 seconds while Taut String took only 12.65 seconds. High efficiency of the proposed denoising method, in addition to its superior performance compared to the other methods, are the main advantages of the preprocessing pipeline. We also observed that using smoothed readcount data decreases the runtime of the CBS segmentation significantly. Therefore, using preprocessing can decrease the overall time complexity of the CNV detection.

V. CONCLUSION

In this study, for having precise CNV detection, we developed an efficient and effective preprocessing pipeline for removing biases and noise from readcount data, generated from WES data. The proposed preprocessing pipeline consists of five blocks: filtering outlier reads, removing GC bias, removing mappability bias, eliminating noise, and normalizing for sample purity. While many CNV detection tools do not use denoising and normalization methods, we showed that employing proper denoising and normalization methods can significantly improve the performance of CNV detection in terms of sensitivity and specificity. We also showed that denoising block plays the most important role in improving the performance of CNV detection.

Based on the characteristics of the readcount data and CNV segments, we developed an efficient non-linear denoising method that can preserve edges and focal alterations. The proposed denoising method is based on the Taut String approach that is an efficient non-linear method from the signal processing field. To evaluate the performance of the Taut String denoising method, we compared the sensitivities in detecting true CNVs and their breakpoints of the CBS segmentation while using no denoising, Taut String, DWT, and MA methods. DWT and MA denoising methods have been used widely in bioinformatics applications. However, this comparison showed that Taut String outperforms DWT and MA in both efficiency and accuracy. Another advantage of using Taut String denoising approach and having smoother signal is that the segmentation method can be run faster, which can decrease the overall complexity of the CNV detection.

To conclude this study, we can say that preprocessing readcount data is essential in precise detection of CNVs; and advanced normalization and noise cancellation methods from other fields, such as signal processing and statistics can be utilized for having effective and efficient preprocessing.

ACKNOWLEDGEMENT

This study was supported by a grant from the National Institutes of Health (NIH, R00LM011595, PI: Nabavi).

Biography

Fatima Zare is a research assistant at the Computer Science and Engineering department of University of Connecticut. She holds a M.Sc. and a B.Sc. degree in Electrical and Computer Engineering. She is currently working on developing novel CNV detection and visualization methods based on statistical signal processing techniques to identify CNVs more accurately and efficiently. Fatima's research interests are signal processing, bioinformatics, machine learning and data mining.

Sardar Ansari is a research fellow in the University of Michigan Department of Computational Medicine and Bioinformatics and a member of Michigan Center for Integrative Research in Clinical Care (MCIRCC). He earned his Ph.D. in Computer Science and his M.S. degree in Statistics at Virginia Commonwealth University in 2013. He received his bachelors degree in Software Engineering from University of Tehran, Electrical and Computer Engineering Department in 2008, and his M.S. in Computer Science from VCU in 2010. His research interests are medical signal and image processing, machine learning, data mining and development of medical devices as well as non-linear and discrete optimization and queuing theory.

Kayvan Najarian is an Associate Professor in the departments of Computational Medicine and Bioinformatics, and Emergency Medicine at the University of Michigan. He also serves as the director of the Michigan Center for Integrative Research in Critical Cares Biosignal-Image and Computational (BIC) Core program. Dr. Najarian received his Ph.D. in Electrical and Computer Engineering from University of British Columbia, Canada, M.Sc in Biomedical Engineering from Amirkabir University, Iran, and B.Sc. in Electrical

Engineering from Sharif University, Iran. The focus of Dr. Kayvan Najarian's research is on the design of signal/image processing and machine learning methods to create computerassisted clinical decision support systems that improve patient care. Dr. Najarian serves as the Editor-in-Chief of a journal in the field of Biomedical Engineering as well as the Associate Editor of two journals in the field of biomedical informatics. He is also a member of many editorial boards and has served as a guest editor of special issues for several journals.

Dr. Sheida Nabavi is an Assistant Professor in the department of Computer Science and Engineering and the Institute for Systems Genomics at the University of Connecticut. She received her PhD from Electrical and Computer Engineering Department at Carnegie Mellon University. Then she joined the Center for Biomedical Informatics at Harvard Medical School (HMS) as an NIH funded Research Fellow and received her Master's in Medical Science focused on bioinformatics. The focus of her research is on development of novel computational methods and algorithms, based on advanced statistical machine learning and signal/image processing techniques, to analyze and integrate high-throughput sequencing data for identifying genomic/epigenetic features associated with different phenotypes, especially cancer.

REFERENCES

[1]. Zarrei M, MacDonald JR, Merico D, and Scherer SW, "A copy number variation map of the human genome," Nature Reviews Genetics, vol. 16, no. 3, pp. 172–183, 2015.

[2]. Shlien A and Malkin D, "Copy number variations and cancer," Genome medicine, vol. 1, no. 6, p. 62, 2009. [PubMed: 19566914]

[3]. Zare F, Dow M, Monteleone N, Hosny A, and Nabavi S, "An evaluation of copy number variation detection tools for cancer using whole exome sequencing data," BMC bioinformatics, vol. 18, no. 1, p. 286, 2017. [PubMed: 28569140]

[4]. Metzker ML, "Sequencing technologiesthe next generation," Nature reviews genetics, vol. 11, no. 1, p. 31, 2010.

[5]. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, and Law M, "Comparison of next-generation sequencing systems," BioMed Research International, vol. 2012, 2012.

[6]. Teo SM, Pawitan Y, Ku CS, Chia KS, and Salim A, "Statistical challenges associated with detecting copy number variations with next-generation sequencing," Bioinformatics, vol. 28, no. 21, pp. 2711–2718, 2012. [PubMed: 22942022]

[7]. Duan J, Wan M, Deng H-W, and Wang Y-P, "A sparse model based detection of copy number variations from exome sequencing data," IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 496–505, 2016. [PubMed: 26258935]

[8]. Wang KYH, Nettleton D, "Copy number variation detection using next generation sequencing read counts," BMC Bioinformatics, vol. 15, no. 1, p. 109, 2014. [PubMed: 24731174]

[9]. Yoon S, Xuan Z, Makarov V, Ye K, and Sebat J, "Sensitive and accurate detection of copy number variants using read depth of coverage," Genome research, vol. 19, no. 9, pp. 1586–1592, 2009. [PubMed: 19657104]

[10]. Dohm JC, Lottaz C, Borodina T, and Himmelbauer H, “Substantial biases in ultra-short read data sets from high-throughput dna sequencing,” Nucleic acids research, vol. 36, no. 16, pp. e105–e105, 2008. [PubMed: 18660515]

[11]. Iakovishina D, Janoueix-Lerosey I, Barillot E, Regnier M, and Boeva V, “Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability,” Bioinformatics, vol. 32, no. 7, pp. 984–992, 2016. [PubMed: 26740523]

[12]. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, and Barillot E, “Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data,” Bioinformatics, vol. 28, no. 3, pp. 423–425, 2011. [PubMed: 22155870]

[13]. Benjamini Y and Speed TP, “Summarizing and correcting the gc content bias in high-throughput sequencing,” Nucleic acids research, p. gks001, 2012.

[14]. Benjamin DJ, Cesarini D, van der Loos MJ, Dawes CT, Koellinger PD, Magnusson PK, Chabris CF, Conley D, Laibson D, Johannesson M et al., “The genetic architecture of economic and political preferences,” Proceedings of the National Academy of Sciences, vol. 109, no. 21, pp. 8026–8031, 2012.

[15]. Liao C, Yin A.-h., Peng C.-f., Fu F, Yang J.-x., Li R, Chen Y.-y., Luo D.-h., Zhang Y.-l., Ou Y.-m.et al., “Noninvasive prenatal diagnosis of common aneuploidies by semiconductor sequencing,” Proceedings of the National Academy of Sciences, vol. 111, no. 20, pp. 7415–7420, 2014.

[16]. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, and Gnirke A, “Analyzing and minimizing pcr amplification bias in illumina sequencing libraries,” Genome biology, vol. 12, no. 2, p. R18, 2011. [PubMed: 21338519]

[17]. Gao F and Zhang C-T, “Gc-profile: a web-based tool for visualizing and analyzing the variation of gc content in genomic sequences,” Nucleic acids research, vol. 34, no. suppl 2, pp. W686–W691, 2006. [PubMed: 16845098]

[18]. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, and Berri S, “Correcting for genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data,” Bioinformatics, vol. 28, no. 1, pp. 40–47, 2012. [PubMed: 22039209]

[19]. Teo SM, Pawitan Y, Ku CS, Chia KS, and Salim A, “Statistical challenges associated with detecting copy number variations with next-generation sequencing,” Bioinformatics, vol. 28, no. 21, pp. 2711–2718, 2012. [PubMed: 22942022]

[20]. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK, “Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing,” Genome research, vol. 22, no. 3, pp. 568–576, 2012. [PubMed: 22300766]

[21]. Lai D, Ha G, Shah S, Lai MD, biocViews Sequencing P, and CopyNumberVariation M, “Package hmmcopy,” 2011.

[22]. Pique-Regi R, Ortega A, Tewfik A, and Asgharzadeh S, “Detecting changes in dna copy number: Reviewing signal processing techniques,” IEEE Signal Processing Magazine, vol. 29, no. 1, pp. 98–107, 2012.

[23]. Metzker ML, “Sequencing technologies?the next generation,” Nature reviews genetics, vol. 11, no. 1, p. 31, 2010.

[24]. Amarasinghe KC, Li J, Hunter SM, Ryland GL, Cowin PA, Campbell IG, and Halgamuge SK, “Inferring copy number and genotype in tumour exome data,” BMC genomics, vol. 15, no. 1, p. 732, 2014. [PubMed: 25167919]

[25]. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, and Hochreiter S, “cn. mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate,” Nucleic acids research, p. gks003, 2012.

[26]. Xi R, Hadjipanayis AG, Luquette LJ, Kim T-M, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA et al., “Copy number variation detection in whole-genome sequencing data using the bayesian information criterion,” Proceedings of the National Academy of Sciences, vol. 108, no. 46, pp. E1128–E1136, 2011.

[27]. Stamoulis C and Betensky RA, “A novel signal processing approach for the detection of copy number variations in the human genome,” Bioinformatics, vol. 27, no. 17, pp. 2338–2345, 2011. [PubMed: 21752800]

[28]. Ben-Yaacov E and Eldar YC, “A fast and flexible method for the segmentation of acgh data,” Bioinformatics, vol. 24, no. 16, pp. i139–i145, 2008.

[29]. Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, and Porter P, “Denoising array-based comparative genomic hybridization data using wavelets,” Biostatistics, vol. 6, no. 2, pp. 211–226, 2005. [PubMed: 15772101]

[30]. Yuan Y, Failmezger H, Rueda OM, Ali HR, Gräf S, Chin S-F, Schwarz RF, Curtis C, Dunning MJ, Bardwell H et al., “Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling,” Science translational medicine, vol. 4, no. 157, pp. 157ra143–157ra143, 2012.

[31]. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B et al., “Allele-specific copy number analysis of tumors,” Proceedings of the National Academy of Sciences, vol. 107, no. 39, pp. 16910–16915, 2010.

[32]. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA et al., “Absolute quantification of somatic dna alterations in human cancer,” Nature biotechnology, vol. 30, no. 5, pp. 413–421, 2012.

[33]. Oesper L, Mahmoody A, and Raphael BJ, “Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data,” Genome biology, vol. 14, no. 7, p. R80, 2013. [PubMed: 23895164]

[34]. Olshen AB, Venkatraman E, Lucito R, and Wigler M, “Circular binary segmentation for the analysis of array-based dna copy number data,” Biostatistics, vol. 5, no. 4, pp. 557–572, 2004. [PubMed: 15475419]

[35]. Quinlan AR and Hall IM, “Bedtools: a flexible suite of utilities for comparing genomic features,” Bioinformatics, vol. 26, no. 6, pp. 841–842, 2010. [PubMed: 20110278]

[36]. Cheung M-S, Down TA, Latorre I, and Ahringer J, “Systematic bias in high-throughput sequencing data and its correction by beads,” Nucleic acids research, p. gkr425, 2011.

[37]. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and others, “The sequence alignment/map format and SAMtools,” Bioinformatics, vol. 25, no. 16, pp. 2078–2079, 2009. [PubMed: 19505943]

[38]. Donoho DL, “De-noising by soft-thresholding,” IEEE transactions on information theory, vol. 41, no. 3, pp. 613–627, 1995.

[39]. Kawasaki T, Takai Y, Ikuta T, and Shimizu R, “Wave field restoration using three-dimensional fourier filtering method,” Ultramicroscopy, vol. 90, no. 1, pp. 47–59, 2001. [PubMed: 11794629]

[40]. Donoho DL, “Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data,” in In Proceedings of Symposia in Applied Mathematics. Citeseer, 1993.

[41]. Ana Sovic DS, “Signal decomposition methods for reducing drawbacks of the dwt,” Engineering Review, vol. 32, no. 2, pp. 70–77, 2012.

[42]. Brunk H, Barlow R, Bartholomew D, and Bremner J, “Statistical inference under order restrictions.(the theory and application of isotonic regression),” MISSOURI UNIV COLUMBIA DEPT OF STATISTICS, Tech. Rep, 1972.

[43]. Belle A, Ansari S, Spadafore M, Convertino VA, Ward KR, Derksen H, and Najarian K, “A signal processing approach for detection of hemodynamic instability before decompensation,” PloS one, vol. 11, no. 2, p. e0148544, 2016. [PubMed: 26871715]

[44]. Davies PL and Kovac A, “Local extremes, runs, strings and multiresolution,” Annals of Statistics, pp. 1–48, 2001.

[45]. Mammen E and van de Geer S, “Penalized quasi-likelihood estimation in partial linear models,” The Annals of Statistics, pp. 1014–1035, 1997.

[46]. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK, “Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing,” Genome research, vol. 22, no. 3, pp. 568–576, 2012. [PubMed: 22300766]

[47]. “DNAcopy.” [Online]. Available: http://bioconductor.org/packages/DNAcopy/

[48]. Li H and Durbin R, “Fast and accurate short read alignment with burrows–wheeler transform,” Bioinformatics, vol. 25, no. 14, pp. 1754–1760, 2009. [PubMed: 19451168]

[49]. Quinlan AR and Hall IM, “Bedtools: a flexible suite of utilities for comparing genomic features,” Bioinformatics, vol. 26, no. 6, pp. 841–842, 2010. [PubMed: 20110278]

[50]. “GenomicRanges.” [Online]. Available: http://bioconductor.org/packages/GenomicRanges/

[51]. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, and Carey VJ, “Software for computing and annotating genomic ranges,” PLoS computational biology, vol. 9, no. 8, p. e1003118, 2013. [PubMed: 23950696]

[52]. Pique-Regi R, Ortega A, Tewfik A, and Asgharzadeh S, “Detecting changes in dna copy number: Reviewing signal processing techniques,” IEEE Signal Processing Magazine, vol. 29, no. 1, pp. 98–107, 2012.

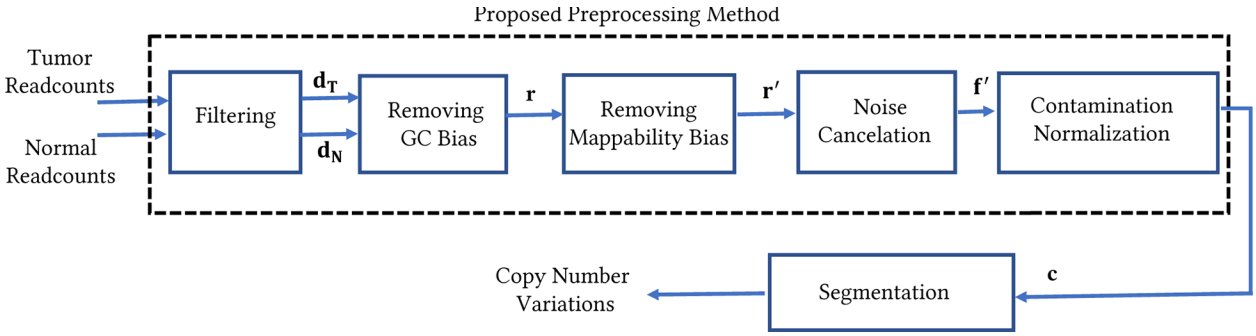

Fig. 1:

The overall copy number variation detection pipeline. Vectors $\mathbf{d_T}$ and $\mathbf{d_N}$ are the number of reads of tumor and normal genome, respectively. The vector r is the *log2ratio* of tumor and normal readcounts after removing GC bias. Vector **r** is the log2 ratio of readcount signal after removing mappability bias. Vector $\mathbf{f}'$ is two to the power of the denoised signal f (copy number ratio). The vector **c** is the normalized signal.

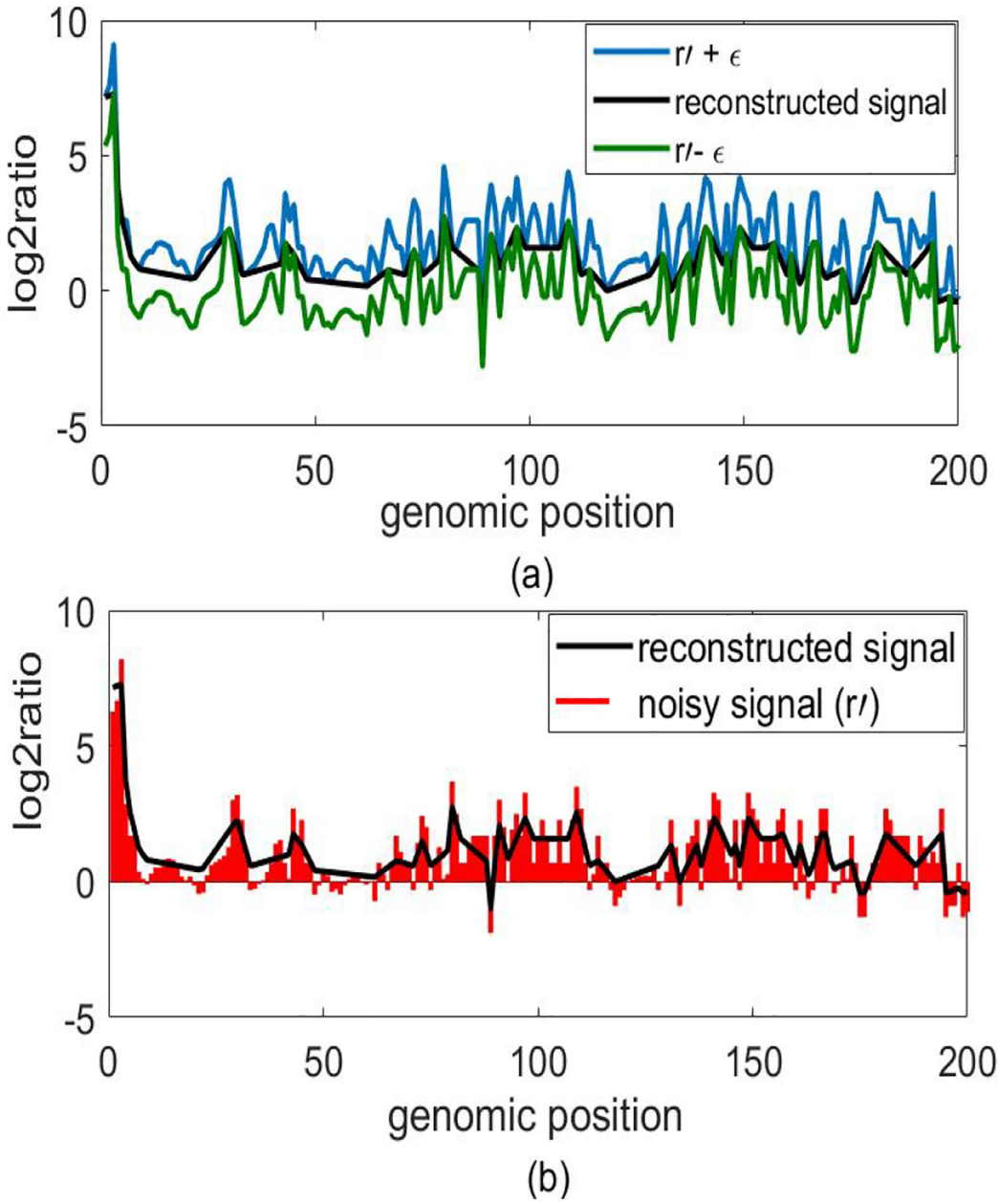

Fig. 2:

Reconstruction of the original readcounts from real breast cancer data using the Taut String smoothing method. ϵ is 0.95. The black line is the estimated smoothed signal.

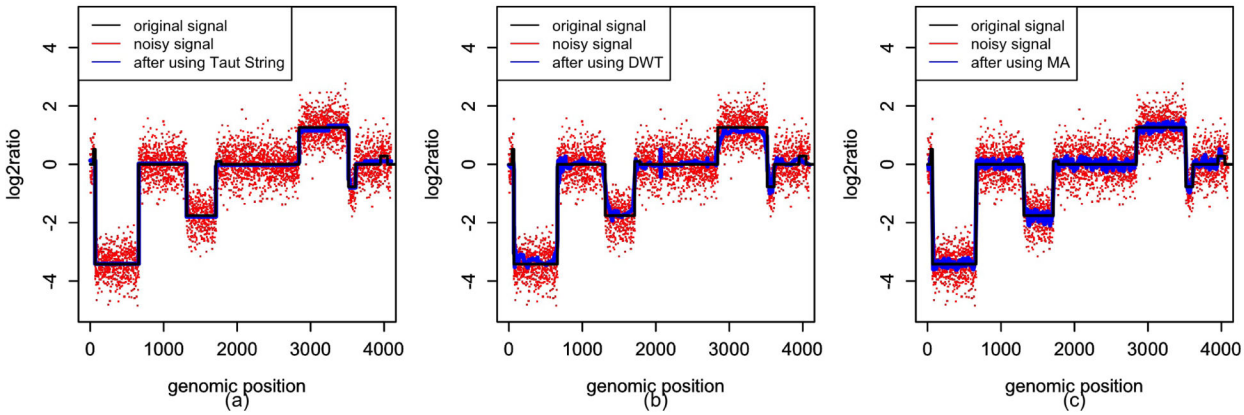

Fig. 3:
Denoising with a) Taut String and b) DWT c) MA. (SNR=7)

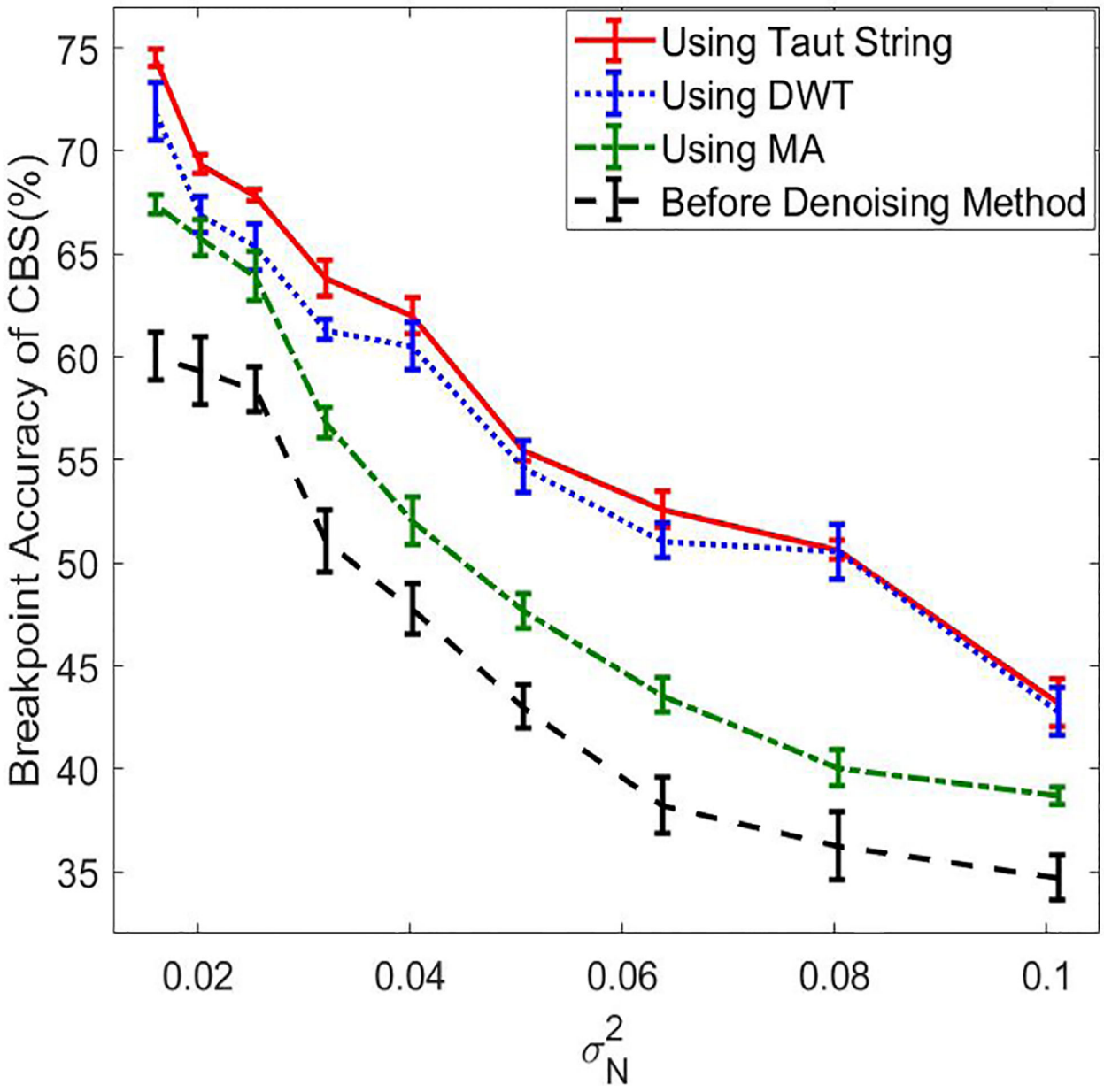

Fig. 4:

Breakpoint accuracy before and after applying denoising for different σ_N^2, using CBS segmentation and simulated readcount data.

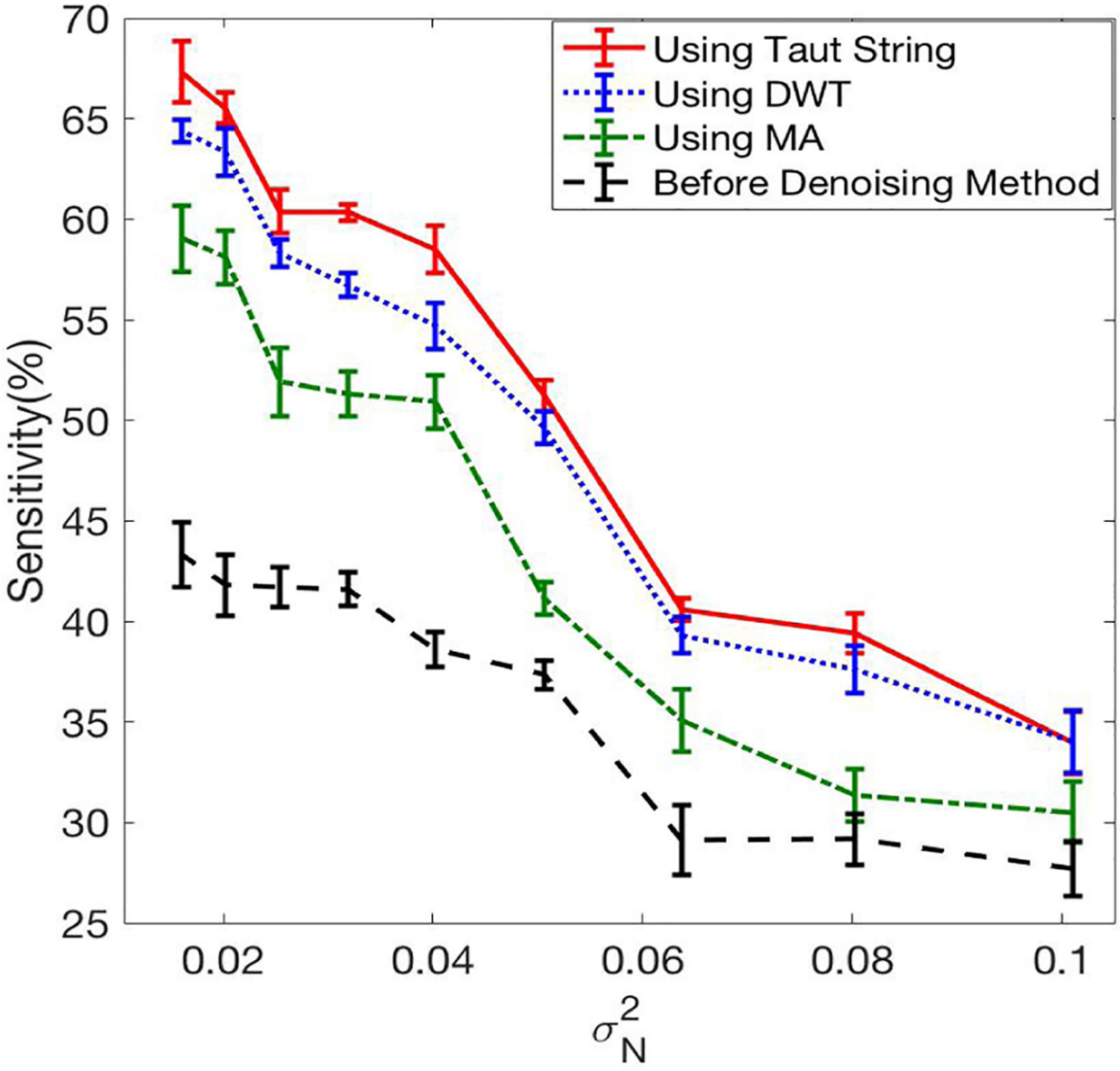

Fig. 5:

Sensitivity of detection of CNVs segments before and after applying denoising methods for different level of noise readcount data. (σ_N^2), using simulated readcount data.

TABLE I:

Possible Outcomes for CNV Genes and Performance Metrics

CNV gene	Not detected	Detected
Present	FN	TP
Not present	TN	FP
Performance metrics: Sensitivity $= \frac{TP}{FN+TP}$	FDR $= \frac{FP}{FP+TP}$	Specificity $= \frac{TN}{FP+TN}$

TABLE II:

Overall Performance of The Denoising Block Using The Simulated WES Data

Preprocessing Steps	Amplification			Deletion		
	Sensitivity	FDR	Specificity	Sensitivity	FDR	Specificity
Before applying preprocessing method	80.53%	31.23%	83.39%	81.97%	33.50%	83.91%
After noise cancelation (Taut String method)	91.39%	22.48%	94.56%	90.78%	20.63%	94.77%
After noise cancelation (DWT method)	91.25%	22.15%	94.40%	91.07%	21.12%	93.95%

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2020 June 08.

TABLE III:

Overall Performance of The Preprocessing Steps Using The Real WES Data

Preprocessing Steps	Amplification			Deletion		
	Sensitivity	FDR	Specificity	Sensitivity	FDR	Specificity
Before applying preprocessing steps	50.99%	42.06%	80.45%	60.37%	64.32%	56.71%
After GC and Mappability correction	58.57%	43.37%	82.69%	68.74%	68.51%	59.20%
After noise cancelation (Taut String method)	70.15%	40.89%	83.45%	80.04%	52.45%	78.45%
After noise cancelation (DWT method)	68.81%	41.65%	79.92%	77.65%	54.32%	72.23%
After tumor contamination normalization using Taut String	72.75%	40.47%	86.39%	84.30%	50.10%	80.21%

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2020 June 08.

TABLE IV:

Overall Performance of The Proposed Preprocessing Pipeline and VarScan2 Using the Real WES Data

Method	Amplification			Deletion		
	Sensitivity	FDR	Specificity	Sensitivity	FDR	Specificity
Varscan2	68.1%	**38.01%**	73.25%	74.31%	56.62%	**80.33%**
Our Proposed Method	**72.75%**	40.47%	**86.39%**	**84.30%**	**50.10%**	80.21%

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2020 June 08.